Linear Regression

Dr. Jarad Niemi

STAT 4610X - Iowa State University

February 6, 2025

Outline

- Simple Linear Regression (SLR)
  - Model
  - Interpretation
  - Assumptions
  - Diagnostics
  - Example
  - Two-sample T-test
- Multiple Linear Regression (MLR)
  - Model
  - Interpretation
  - Assumptions
  - Diagnostics
  - Examples

Simple Linear Regression

For observation $i = \{1, 2, \ldots, \}$, let

- Y_i be the response variable and
- X_i be the explanatory variable.

The simple linear regression model (SLR) assumes

$$Y_i \stackrel{ind}{\sim} N(\beta_0 + \beta_1 X_i, \sigma^2)$$

or, equivalently,

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad \epsilon_i \stackrel{ind}{\sim} N(0, \sigma^2).$$

Interpretation

Recall

$$E[Y_i] = \beta_0 + \beta_1 X_i$$

Thus,

- β_0 is the expected response when $X_i = 0$
- β_1 is the expected increase in the response when X_i is increased by 1.

Assumptions

Recall

$$E[Y_i] = \beta_0 + \beta_1 X_i, \quad \epsilon_i \stackrel{ind}{\sim} N(0, \sigma^2)$$

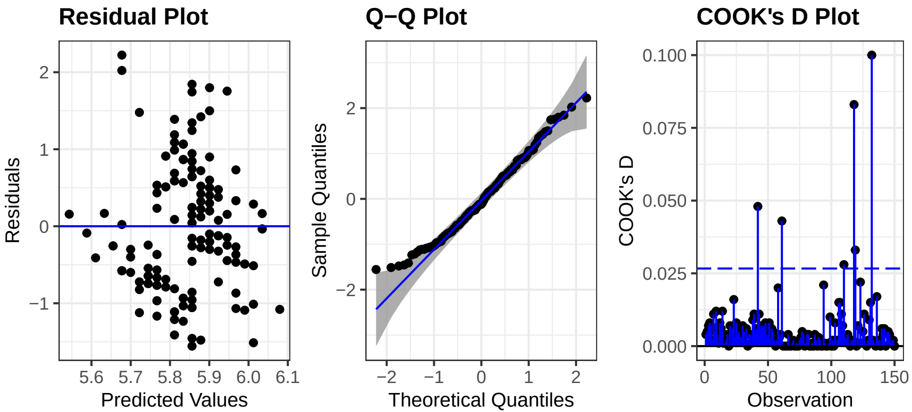
Thus, the model assumptions are

- The errors are independent.
- The errors are normally distributed.
- The errors have constant variance.
- The relationship between the expected response and the explanatory variable is a straight line.

Diagnostics

To evaluate these model assumptions we utilize diagnostic plots:

```
m <- lm(Sepal.Length ~ Sepal.Width, data = iris)
ggResidpanel::resid_panel(m, plots = c("resid", "qq", "cookd"), qqbands = TRUE, nrow = 1)
```



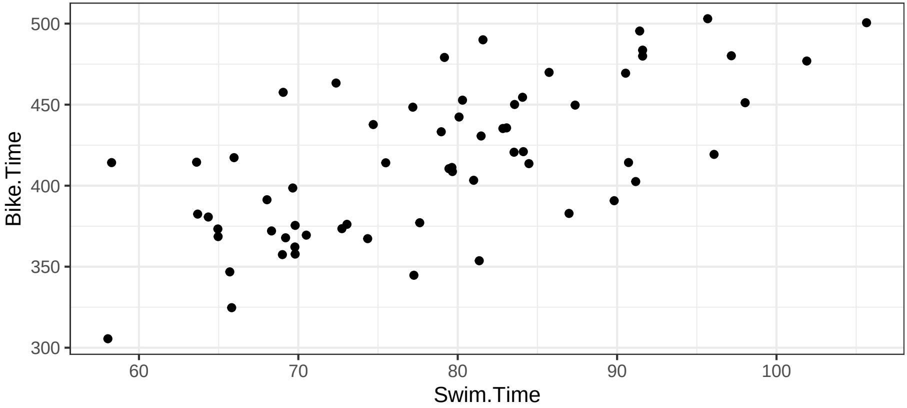
Triathlon Data

from https://modules.scorenetwork.org/triathlons/ironman-lakeplacid-mlr/

```
d <- read_csv("ironman_lake_placid_female_2022_canadian.csv")
head(d)

# A tibble: 6 x 17
    Bib Name     Country Gender Division Division.Rank Overall.Time Overall.Rank Swim.Time Swim.Rank Bike
  <dbl> <chr>    <chr>   <chr>  <chr>            <dbl>        <dbl>        <dbl>     <dbl>     <dbl>
1     2 Melanie~ Canada  Female FPRO                 5         575.           21      58.0        57
2     9 Pamela-~ Canada  Female FPRO                10         610.           51      65.8       253
3  1000 Carley ~ Canada  Female F35-39               4         660.          126      65.7       249
4  1935 Seanna ~ Canada  Female F45-49               3         665.          131      74.4       727
5   511 Marie-C~ Canada  Female F45-49               4         679.          161      77.2       899
6  1240 Julie H~ Canada  Female F40-44               6         693.          202      77.6       921
# i 5 more variables: Run.Time <dbl>, Run.Rank <dbl>, Finish.Status <chr>, Location <chr>, Year <dbl>
```

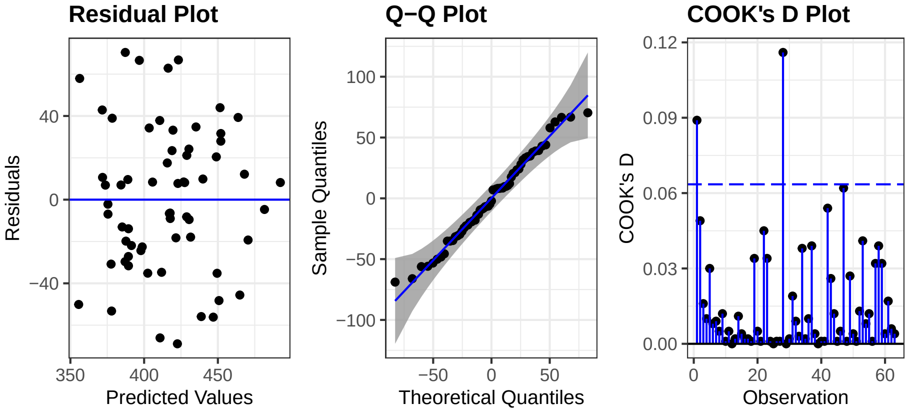
Bike Time v Swim Time

```
ggplot(d |> filter(Swim.Time < 500), aes(x = Swim.Time, y = Bike.Time)) + geom_point()
```



Bike Time v Swim Time - Model Diagnostics

```
m <- lm(Bike.Time ~ Swim.Time, data = d |> filter(Swim.Time < 500))
ggResidpanel::resid_panel(m, plots = c("resid", "qq", "cookd"), qqbands = TRUE, nrow = 1)
```



Bike Time v Swim Time - Model Results

```
summary(m)



Call:
lm(formula = Bike.Time ~ Swim.Time, data = filter(d, Swim.Time <
    500))

Residuals:
    Min      1Q  Median      3Q     Max
-68.901 -23.468  -2.169  23.808  70.369

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 188.8604    32.0893   5.885 1.82e-07 ***
Swim.Time     2.8729     0.4035   7.120 1.44e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 34.83 on 61 degrees of freedom
Multiple R-squared:  0.4538,Adjusted R-squared:  0.4449
F-statistic: 50.69 on 1 and 61 DF,  p-value: 1.443e-09
```

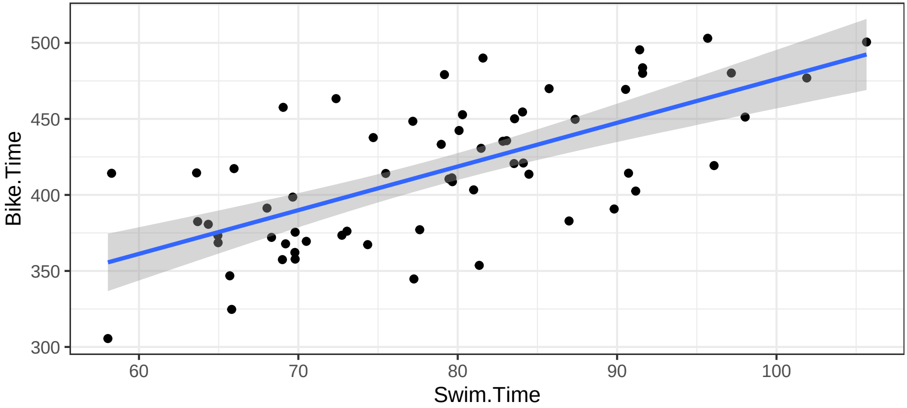
Bike Time v Swim Time - Written Results

```
cbind(coef(m), confint(m))

                                2.5 %     97.5 %
(Intercept) 188.860386 124.693942 253.026829
Swim.Time     2.872855   2.065987   3.679724

summary(m)$r.squared

[1] 0.4538433
```

When swim time is 0, the expected Bike Time is 189 mins with a 95% interval of (125, 253).
For additional minute of swim time, the bike time is expected to increase 2.9 mins (2.1, 3.7).
The model explains 45% of the variability in bike time.

Bike Time v Swim Time - Plot

```
ggplot(d |> filter(Swim.Time < 500), aes(x = Swim.Time, y = Bike.Time)) +
  geom_point() + geom_smooth(method = "lm")
```



Comparing two groups

We can use SLR to compare two groups. Note that

$$Y_i \stackrel{ind}{\sim} N(\mu_{g[i]}, \sigma^2)$$

where $g[i] \in \{1, 2\}$ determines the group membership for observation i is equivalent to

$$Y_i \stackrel{ind}{\sim} N(\beta_0 + \beta_1 \mathrm{I}(g[i] = 2), \sigma^2)$$

where $\mathrm{I}(g[i] = 2)$ is the indicator function, i.e.

$$I(A) = \begin{cases} 1 & A \text{ is TRUE} \\ 0 & \text{otherwise} \end{cases}$$

and

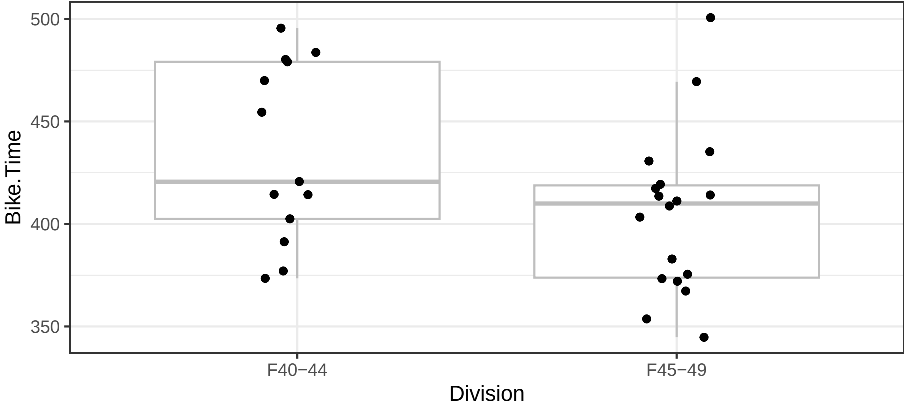
$$\mu_1 = \beta_0 \quad \text{and} \quad \mu_2 = \beta_0 + \beta_1.$$

Comparing Bike Times for Two Age Divisions

```
d2 <- d |> filter(Division %in% c("F40-44", "F45-49"))

d2 |>
  group_by(Division) |>
  summarize(
    n = n(),
    mean = mean(Bike.Time),
    sd = sd(Bike.Time)
  )

# A tibble: 2 x 4
  Division     n  mean    sd
  <chr>    <int> <dbl> <dbl>
1 F40-44      13  435.  43.6
2 F45-49      18  405.  39.5
```

Plotting Bike Times for Two Age Divisions

```
ggplot(d2, aes(x = Division, y = Bike.Time)) +
  geom_boxplot(outliers = FALSE, color = "gray") + geom_jitter(width = 0.1)
```



Modeling Bike Time by Two Age Divisions

```
m <- lm(Bike.Time ~ Division, data = d2)
summary(m)



Call:
lm(formula = Bike.Time ~ Division, data = d2)

Residuals:
    Min      1Q  Median      3Q     Max 
-61.663 -32.237   3.556  27.806  95.406 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)      435.13      11.44  38.040   <2e-16 ***
DivisionF45-49   -29.97      15.01  -1.996   0.0554 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 41.24 on 29 degrees of freedom
Multiple R-squared:  0.1208,Adjusted R-squared:  0.09051 
F-statistic: 3.986 on 1 and 29 DF,  p-value: 0.05535
```

Two-sample T-test

```
cbind(coef(m), confint(m))

                              2.5 %      97.5 %
(Intercept)    435.1295 411.73435 458.5246236
DivisionF45-49 -29.9693 -60.67155   0.7329461

t.test(Bike.Time ~ Division, data = d2, var.equal = TRUE)



Two Sample t-test

data:  Bike.Time by Division
t = 1.9964, df = 29, p-value = 0.05535
alternative hypothesis: true difference in means between group F40-44 and group F45-49 is not equal to 0
95 percent confidence interval:
 -0.7329461 60.6715501
sample estimates:
mean in group F40-44 mean in group F45-49
            435.1295             405.1602
```

(Multiple Linear) Regression

For observation $i = \{1, 2, \dots, n\}$, let

- Y_i be the value of the response variable and
- $X_{i,j}$ be value of the jth explanatory variable

The (multiple linear) regression model assumes

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \cdots + \beta_p X_{i,p} + \epsilon_i$$

and

$$\epsilon_i \stackrel{ind}{\sim} N(0, \sigma^2).$$

Interpretation

Recall

$$E[Y_i] = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \cdots + \beta_p X_{i,p}$$

Thus,

- β_0 is the expected response when all $X_{i,j} = 0$
- β_j is the expected increase in the response when $X_{i,j}$ is increased by 1 and all other explanatory variables are held constant

When multiple regression is used, you will often see people write the phrases "after controlling for" or "after adjusting for" followed by a list of the other explanatory variables in the model.

Assumptions

Recall

$$E[Y_i] = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \cdots + \beta_p X_{i,p}, \quad \epsilon_i \stackrel{ind}{\sim} N(0, \sigma^2)$$

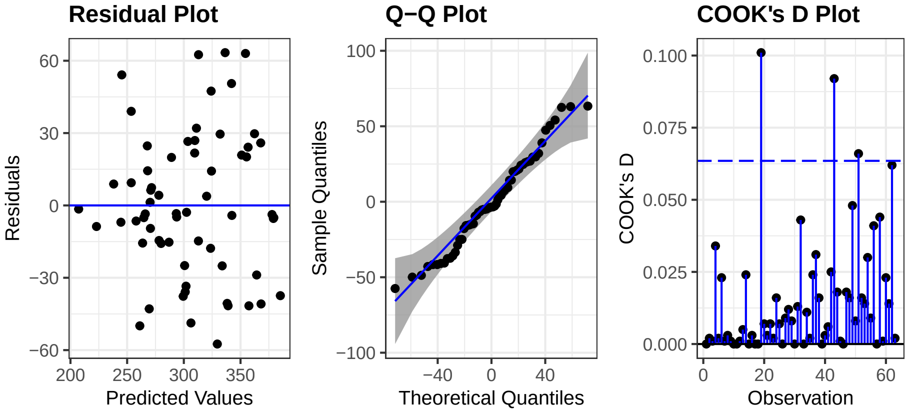
Thus, the model assumptions are

- The errors are independent.
- The errors are normally distributed.
- The errors have constant variance.
- The relationship between the expected response and the explanatory variables is given above.

Diagnostics

To evaluate these model assumptions we utilize diagnostic plots:

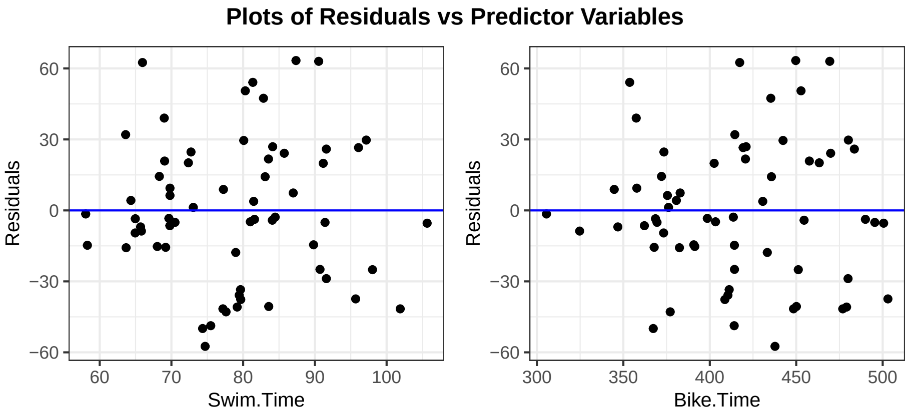
```
m <- lm(Run.Time ~ Swim.Time + Bike.Time, data = d |> filter(Swim.Time < 500))
ggResidpanel::resid_panel(m, plots = c("resid", "qq", "cookd"), qqbands = TRUE, nrow = 1)
```



Diagnostics

To evaluate the need for a quadratic term:

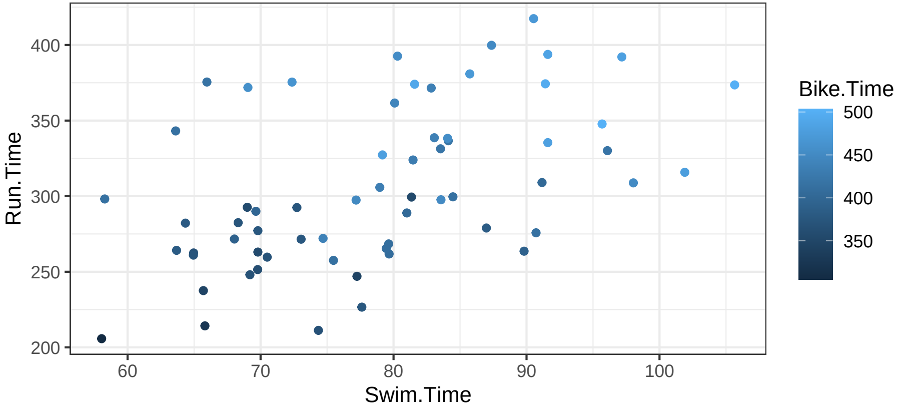
```
m <- lm(Run.Time ~ Swim.Time + Bike.Time, data = d |> filter(Swim.Time < 500))
ggResidpanel::resid_xpanel(m)
```



```
summary(m)


Call:
lm(formula = Run.Time ~ Swim.Time + Bike.Time, data = filter(d,
    Swim.Time < 500))

Residuals:
    Min      1Q  Median      3Q     Max
-57.474 -16.782  -3.523  21.298  63.349

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -68.0058    35.1258  -1.936   0.0576 .
Swim.Time    -0.3771     0.4773  -0.790   0.4326
Bike.Time     0.9726     0.1119   8.689  3.3e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 30.45 on 60 degrees of freedom
Multiple R-squared:  0.6711,Adjusted R-squared:  0.6602
F-statistic: 61.22 on 2 and 60 DF,  p-value: 3.238e-15
```

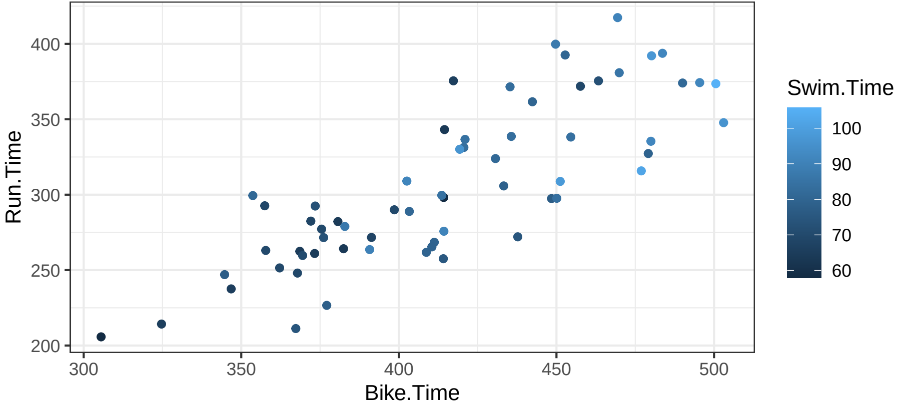
Run Time Plots

```
ggplot(d |> filter(Swim.Time < 500),
       aes(x = Swim.Time, y = Run.Time, color = Bike.Time)) +
  geom_point()
```



Run Time Plots

```
ggplot(d |> filter(Swim.Time < 500),
       aes(x = Bike.Time, y = Run.Time, color = Swim.Time)) +
  geom_point()
```



Written Summary

```
cbind(coef(m), confint(m))
```

```
                                2.5 %   97.5 %
(Intercept) -68.0057881 -138.267938 2.256362
Swim.Time    -0.3771479   -1.331933 0.577637
Bike.Time     0.9725912    0.748696 1.196486
```

```
summary(m)$r.squared
```

```
[1] 0.6711405
```

Using the 2022 Women's Lake Placid Ironman data, we fit a regression model using run time as the response variable and swim and bike times as the explanatory variables. After adjusting for bike time, each minute increase of swim time was associated with a -0.38 minute increase in run time with a 95% interval of (-1.33, 0.58). After adjusting for swim time, each minute increase of bike time was associated with a 0.97 (0.75, 1.2) minute increase in run time. The model with swim and bike time accounted for 67% of the variability in run time.

ANOVA

When our explanatory variable is categorical with more than 2 levels, we can fit a regression model that will often be referred to as an ANOVA model.

To fit this model, we do the following

1. Choose one level to be the reference level (by default R will choose the level that comes first alphabetically)
2. Create indicator variables for all the other levels, i.e.

$$\mathrm{I}(\text{level for observation } i \text{ is } <\!\text{level}\!>) = \begin{cases} 1 & \text{if level for observation } i \text{ is } <\!\text{level}\!> \\ 0 & \text{otherwise} \end{cases}$$

3. Fit a regression model using these indicators.

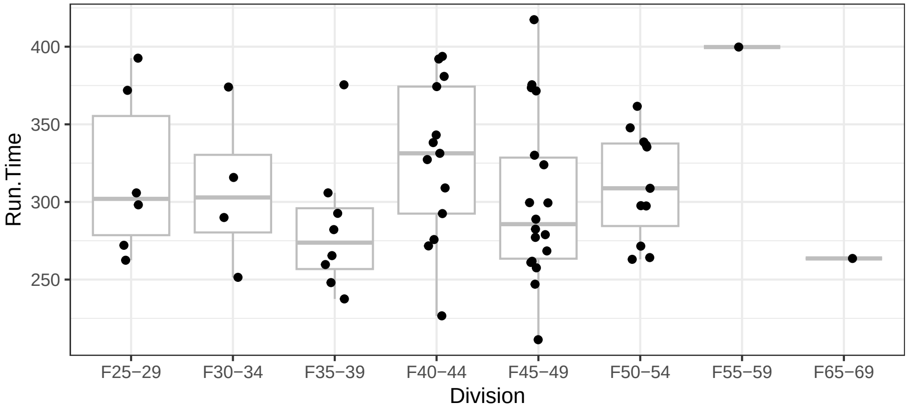
Most statistical software will perform these actions for you, but it is useful to know this is what is happening.

Run Time by Age Group

```
d |> group_by(Division) |>
  summarize(
    n    = n(),
    mean = mean(Run.Time),
    sd   = sd(Run.Time)
  )

# A tibble: 9 x 4
  Division     n  mean    sd
  <chr>    <int> <dbl> <dbl>
1 F25-29       6  317.  53.3
2 F30-34       4  308.  51.5
3 F35-39       8  283.  43.6
4 F40-44      13  327.  51.2
5 F45-49      18  301.  53.9
6 F50-54      11  311.  35.1
7 F55-59       1  400.  NA
8 F65-69       1  264.  NA
9 FPRO         2  210.   6.00
```

Run Time by Division

```r
ggplot(d |> filter(Division != "FPRO"),
       aes(x = Division, y = Run.Time)) +
  geom_boxplot(outliers = FALSE, color = "gray") +
  geom_jitter(width = 0.1)
```



```
m <- lm(Run.Time ~ Division, data = d |> filter(Division != "FPRO"))
summary(m)


Call:
lm(formula = Run.Time ~ Division, data = filter(d, Division !=
    "FPRO"))

Residuals:
     Min       1Q   Median       3Q      Max
-100.808  -35.221   -2.173   26.991  115.983

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     317.164     19.924  15.918   <2e-16 ***
DivisionF30-34   -9.351     31.503  -0.297    0.768
DivisionF35-39  -33.816     26.358  -1.283    0.205
DivisionF40-44   10.260     24.087   0.426    0.672
DivisionF45-49  -15.747     23.007  -0.684    0.497
DivisionF50-54   -6.017     24.769  -0.243    0.809
DivisionF55-59   82.619     52.715   1.567    0.123
DivisionF65-69  -53.564     52.715  -1.016    0.314
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

F-test

When evaluating the statistical support for including a categorical variable with more than 2 levels, we use an F-test. If this variable is the only variable in the model, we can use

```
anova(m)

Analysis of Variance Table

Response: Run.Time
          Df Sum Sq Mean Sq F value Pr(>F)
Division   7  21469  3067.1  1.2877  0.274
Residuals 54 128622  2381.9

drop1(m, test = "F")

Single term deletions

Model:
Run.Time ~ Division
         Df Sum of Sq    RSS    AIC F value Pr(>F)
<none>                128622 489.52
Division  7     21469 150091 485.10  1.2877  0.274
```

ANOVA F-test

Alternatively, and more generally, we can fit a model with and without the variable of interest and compare the two models:

```
m0 <- lm(Run.Time ~ 1, data = d |> filter(Division != "FPRO"))
anova(m0, m)

Analysis of Variance Table

Model 1: Run.Time ~ 1
Model 2: Run.Time ~ Division
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     61 150091
2     54 128622  7     21469 1.2877  0.274
```

Interpretation

```
cbind(coef(m)[c(1,3)], confint(m)[c(1,3),]) # divide by 60 to get hours

                                2.5 %    97.5 %
(Intercept)    317.16389 277.2179 357.10992
DivisionF35-39 -33.81597 -86.6596  19.02766

summary(m)$r.squared

[1] 0.1430418
```

Using the 2022 Women's Lake Placid Ironman data, we fit a regression model using run time as the response variable and age division as the explanatory variable. The mean run time for the F25-29 division was 5.3 hours with a 95% interval of (4.6, 6). The estimated difference in run time for the F25-29 division minus the F35-39 division was 34 (-19, 87) minutes. The model with division accounted for 14% of the variability in run time.